JPL PUBLICATION 81-50

User Requirements for NASA Data Base Management Systems

Part One: Oceanographic Discipline

Brad Fujimoto

(NASA-CR-164701) USER REQUIREMENTS FOR NASA DATA BASE MANAGEMENT SYSTEMS. PART 1: OCEANOGRAPHIC DISCIPLINE (Jet Propulsion Lab.) 58 p HC A04/MF A01 CSCL 05B N81-31019 Unclas G3/82 27245

June 15, 1981

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL PUBLICATION 81-50

User Requirements for NASA Data Base Management Systems

Part One: Oceanographic Discipline

Brad Fujimoto

June 15, 1981

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

ABSTRACT

Generic oceanographic user requirements have been collected and analyzed. These requirements will be used to develop a general multi-purpose data base management system for future missions of the Office of Space and Terrestrial Applications (OSTA) of NASA.

The collection of user requirements involved several separate activities: studying the state-of-the-art technology in data base management systems, analyzing the results of related studies, formulating a viable and diverse list of scientists to be interviewed, developing a presentation format and materials, and interviewing oceanographic data users. In the future, it will be necessary to develop and implement more effective data management systems to handle the increasing influx of data.

ACKNOWLEDGMENTS

The author thanks James Marsh and Vincent Grano, Goddard Space Flight Center (GSFC); Warren Yogi and Rodger Langland, Ocean Data Systems, Inc., (ODSI); and Michael Parke, Jet Propulsion Laboratory (JPL) for contributing their valuable time and cooperating with this study. Appreciation is also expressed to Alfonso Cardenas, consultant for the Applications Data Base Management System (ADBMS) task, for his advice in preparing the methodology for the user requirements task, and to David Lame for his continued support of the same task.

PREFACE

This user requirements study was conducted as part of the Applications Data Base Management System (ADBMS) Project at JPL. It was reviewed by the oceanographic data users interviewed at JPL; GSFC; and ODSI, the ADBMS and Oceanic Pilot System (OPS) Project staff, and the Information Systems Program Office. This study presents a specific solution with a generic point of view to the broader problem of data management. High-level general user requirements for OPS are detailed in another JPL document, Oceanic Pilot System User Functional Requirements, by D. B. Lame (to be published).

CONTENTS

EXECUTIVE SUMMARY

There is an immediate need to develop a data management capability that will help the user community to easily ascertain what space- and surface-acquired information exists on their research and operational activities. Mechanisms must be provided so that multi-source/multi-parameter information can be easily retrieved and used. This report presents user requirements for such an Applications Data Base Management System (ADBMS). This phase of the ADBMS Project addresses user requirements for an oceanographic data management system. These requirements were collected from user interviews, the Seasat Commercial Demonstration Program (Refs. 3, 11), and applicable NASA documents. Relevant requirements were also incorporated from studies in other disciplines. Detailed outlines of the user interviews are provided in Section 4 and a list of references is in Section 5.

The ADBMS should be capable of:

(1) Data types, selection, and delivery

(a) Handling global/local data bases, surface and/or satellite data, near-real-time and/or archived data

(b) Multiple key access, particularly time, geographic location, platform, sensor (for example, altimeter), parameter (for example, wind speed), and resolution

(c) Data delivery via terminal (graphic, tabular), direct computer-to-computer transmission, and mail (plot, photo negative, magnetic tape)

(d) Handling on-line interactive access and batch

(2) User interface

(a) User-friendly command language (logical, English-like), compatible with user programming languages

(b) Flexibility and compatibility

(c) Data independence (user and user's application programs not required to know physical location or format of the data)

(d) Data shareability and nonredundancy, providing efficient, simultaneous access to the same data in a multiple user environment

(3) Data manipulation

(a) Conditional search and retrieval

(b) Reformatting

(c) Quality control

(d) Data selectivity, removal of extraneous data

(e) Cross-referencing of files

(f) Comparing/sorting/merging data sets

(g) Graphics capabilities, overlaying

(4) Documentation

(a) An on-line catalog that incorporates pointers to the entire data base and permits interactive browsing using multiple keys to locate desired data

(b) Complete information data base including source, coverage, quality assessment, related data sets, algorithms, availability, scientific contact, references, processing time and cost, project software documentation, and user application programs

(5) Other System Capabilities

(a) Security

(b) Private file capability

(c) Product archive capability

(d) Application program capability

(e) Report writer capability

(6) System performance

(a) Fast response times

(b) Efficient processing

(c) High availability

(d) Maintain data integrity

(e) Data network potential

(f) Near-real-time capability

For the previous categories, user requirements cannot be simply identified for the entire oceanographic community. The oceanographic data users have too wide a range of: (1) data requirements (e.g., satellite/surface data, near-real-time/archived data, etc.), (2) performance requirements (e.g., timeliness,

response times, availability), (3) research applications, (4) processing facilities (e.g., computer time), (5) software, and (6) interface requirements.

Users are concerned about the significant problems in data delivery performance, data handling, and data utilization. The key areas in which improvements are needed are in the reduction of costs and processing time, faster data delivery, and elimination of the need to maintain large tape libraries.

Cost reductions could be effected in part by more efficient data manipulation techniques and tools. Programmers are spending large proportions of their time manipulating data sets to facilitate analysis and to make the data compatible for application programs. For many users, the costs outweigh the utility. Many academic users may be prevented from working with existing data sets due to lack of sufficient funds, facilities, and manpower.

Improved product selection techniques and data delivery modes also would reduce costs. Usually, programmers have to format and reduce data sets. Multi-key selection is one means that would provide the user with more extensive control over data selection criteria (e.g., time, location, sensor, parameter). Multiple delivery modes (e.g., terminal, computer-to-computer, high speed printer, magnetic tape) enable the user to select the optimum medium for one's application. On-line, interactive access could significantly lessen costs by reducing the need to read, reformat, store and maintain magnetic tapes.

System performance must be responsive to the needs of the users to encourage utilization. For example, the system must be easy to use, flexible, and compatible with user systems and applications. The capability to adapt to changes in user requirements and data analysis techniques is critical. Data integrity is also critical; the users require good quality data. Documentation must be accessible, comprehensive, and useful to the inexperienced user.

The ADBMS should be capable of handling real-time data. This is not to advocate real-time systems, but rather the capability to implement the ADBMS on these highly interactive systems that demand efficient processing and timely data delivery. Scientists anticipate more real-time applications; some of the techniques and programs that they develop will be used in real-time. Other requirements consist of the capability to use DBMS commands in application programs and the capability to have the ADBMS access data bases on interconnected systems.

Data requirements of the oceanographic community will continue to increase as more data becomes available. The minimum requirements for performance (e.g., data delivery) and convenience (e.g., data medium) become even more critical. If users cannot easily access and use the data, the data are essentially unavailable and worthless. These generic user requirements stress the need to facilitate data utilization and to manage the increasing influx of data.

SECTION 1

INTRODUCTION

1.1 OBJECTIVE

The objective of this document is to define detailed user requirements pertaining to a general-purpose, multi-mission data base management system (DBMS) for future missions of the Office of Space and Terrestrial Applications (OSTA) of NASA. The results of this study will be provided to the Oceanic Pilot System (OPS) Project of JPL.

1.2 STATEMENT OF THE PROBLEM

From the users' perspective, the problems in finding, accessing, and utilizing oceanographic data have been barriers to efficient use. Some users have found that current delivery times preclude utilization; scientists lose valuable time waiting months or years to receive data. Often, there is not sufficient information about a data set, and the appropriate scientific sources are difficult to locate. Many users need to know what has been done to the data or to find information on the format. Users also consume large amounts of processing time to reformat and merge data sets.

The maintenance of a user's magnetic tape library also poses a significant problem. After a tape is received and reformatted, the user must maintain the integrity of the data by storing several copies. The user must archive the tapes as it requires too much time and effort to retrieve the same data from the source at a later time. This activity consumes valuable programmer time and creates significant storage problems. These problems and costs make it extremely difficult, if not impossible, for users with limited resources to use oceanographic data.

It would be more resource and cost efficient to lessen the duplication in software development among the users. A frequently updated software and application information data base could reduce this duplication of effort.

1.3 APPROACH

The information presented in this report has been derived from user interviews and other NASA data system documents and studies. This report, however, should serve only as an initial study because the interviews were limited to oceanographic data users at JPL and Goddard Space Flight Center (GSFC). The initial findings presented will be developed in the future phases of this task, which will also include more interviews with the scientific data users.

The collection of user requirements involved several separate activities: studying the state-of-the-art technology in data base management systems (DBMS), analyzing the results of related studies, formulating a viable and diverse list of scientists to be interviewed, developing a presentation

format and presentation materials, and interviewing oceanographic data users at JPL, GSFC, and Ocean Data Systems, Inc. (ODSI). A DBMS course presented the state of current DBMS technology and assisted the task staff in preparing the methodology for the user interviews. This was necessary as DBMS technology has been rapidly changing and progressing. Current information must be disseminated to the potential users to obtain relevant user requirements. The articulation of user requirements is often biased by what the user thinks is technologically feasible or practical.

The requirements presented in this document are preliminary. For some DBMS functions, no definitive requirements could be established. There are many difficulties in assessing user needs and utilization of the data when it involves the development of a new technology. The requirements cannot be complete, as the potential user community cannot fully grasp the utility of DBMS functions until after implementation. Implementation, however, is only possible after a prototype DBMS has been designed and developed. Possibly after the users have had some experience with a DBMS, they will be able to describe their needs more completely. At present, the only feasible action is to assess user requirements through a broad survey of oceanographic data users and to make and indicate educated suppositions where possible.

The list of oceanographic data users was developed through JPL contacts in the Seasat Project, the Ocean Data Utilization System (ODUS) Project, the Oceanic Pilot System (OPS) Project, and through the DBMS consultant on this task. To facilitate the study, most of the users selected have data processing experience. This experience served to simplify the interviews because the scientists were familiar with computer hardware/software and with the data processing procedures used in their research activities. The list included users with diverse data requirements and activities; this would result in a more balanced and representative set of requirements. This diversity included the data types utilized (for example, satellite or surface data), processing capability, affiliation (for example, government, academic, or commercial), and data sources. The list was severely reduced by the limitation to JPL and GSFC. As a result, information obtained through previous contacts was used, particularly the Seasat Commercial Demonstration Program.

Commercial marine requirements have been incorporated in this document and derived from the Seasat Commercial Demonstration Program. A joint NASA/U.S. Navy marine weather data distribution system, the Seasat Satellite Data Distribution System (SDDS), is located at the Fleet Numerical Oceanography Center (FNOC) in Monterey, California. The SDDS provides a set of current meteorological and oceanographic data products every 12 hours, prepared by FNOC, to members of the commercial marine community. Information on the SDDS commercial users pertaining to DBMS user requirements has been incorporated and distinguished from the academic community, where applicable. For comparison, significant differences between operational and research environments are detailed in Sections 3.1.1 and 3.3.1; however, there are many common requirements between the two groups. Many SDDS users use data for both operational and research activities.

The interview process was preceded by the development of presentation format and materials. This consisted of a detailed, general-application

outline and a "strawman." The outline served as a guide; the interviews did not necessarily follow the logical order of the outline. Some areas were discussed in depth, whereas inapplicable topics were skipped.

The "strawman" was designed from an actual oceanographic data study to indicate the data requisition, delivery, and processing procedures for a sample data user. This example was drawn for two scenarios: a realistic current situation and a possible future situation using an advanced DBMS on a data network. The data catalog example from the NEEDS functional requirements document (Ref. 4) was incorporated. The outline for user interviews is in Appendix A and the strawman in Appendixes B-C.

Summaries of the user interviews conducted at GSFC, JPL, and ODSI are provided in Section 4. A list of the NASA documents incorporated in this report is provided in the references section.

SECTION 2

USER ANALYSIS

2.1 USER ACTIVITIES

The potential user community for oceanographic data from future NASA systems is composed of government, academic, and commercial users. The activities for these groups are divided into two categories: research and operations.

In an operational mode, a user is usually using the data for everyday or time-dependent functions (for example, weather forecasting and ship routing). This requires near-real-time (NRT) data that is passed on to the end user (for example, a vessel) as soon as possible; the value of the data decreases with age (Refs. 11, 14).

In contrast, research activities usually analyze data for periods longer than six months as compared to hours in an operational setting. As a result, the data collection time or timeliness is not critical for most research applications (for example, Seasat data studies involve satellite data collected during mid-1978) (Refs. 8, 10).

The scope of activities within each user group is quite broad. For example, some of the activities in the government sector include tide and current studies (Ref. 10), marine safety (Coast Guard), and weather forecasting (NWS). For this reason, user requirements in this report are based on user characteristics rather than on general user groups. Some of these characteristics are frequent/infrequent data users and near-real-time/archived data users. A frequent data user is defined as one who makes frequent data requests (for example, weekly); whereas, an infrequent data user may need only a few data sets for a long-term project. A frequent data user will place more stringent requirements on the processing performance and interface capabilities of a DBMS (see Sections 3.1.1 and 3.6.1). Alternatively, an infrequent data user will need documentation oriented toward the first-time user and simplicity in operation (see Sections 3.2.1 and 3.4).

In the interviews, the activities of the government users consisted of oceanographic research on tides, currents, topography, and geodetics. The users interviewed also foresaw near-real-time applications, the development of programs to be used on a real-time system and the identification and study of real-time oceanographic phenomena (Refs. 8, 10).

The activities of the SDDS commercial users consist of offshore oil and gas exploration, ocean mining, ocean routing, fishing and environmental forecasting. Most of these activities require near-real-time data; however, longer-term, non-real-time applications exist. For example, an oil company may want the 20-year history of high winds and wave heights for a future drilling site (Ref. 11).

2.2 DATA REQUIREMENTS

The volume of oceanographic data is rapidly expanding, and an increasing number of users want access to this data base. The term "data base" in this context is actually a mixture of different sets of data, physically and logically distributed among several agencies.

For the government and academic sectors, data requirements vary among the different activities and from one request to the next. In research, the approach to a problem usually is not standardized. The investigator may vary the data types, data sources, time periods, and regions, but the selection depends on the quality, availability, and the analysis or application.

In contrast, the commercial user usually makes recurring data requests. For example, a weather forecasting company receives a global data set every 12 hours from the same data supplier (Ref. 14). The data requirements of the commercial user are usually simple, structured, and may be specified in advance of the need. Commercial users with different activities, processing routines, and products may have identical data requirements (for example, weather forecasting and ship routing require global data sets) (Refs. 7, 11).

2.3 COMPUTER FACILITIES

All of the users interviewed have access to large computer systems with disk and tape storage (for details, see Section 4). This is not representative of the entire oceanographic community; some users in the oceanographic community may not have sufficient funding, computer processing time, or computer facilities. The users interviewed indicated that they prefer to perform the necessary data analyses on their own systems. Data users with few processing resources, particularly money, may lean on NASA to provide these services (for example, processing time); many academic users may fit into this category (Refs. 8, 10, 14).

The most significant problem in this area is the maintenance and handling of large tape libraries. At present, the most common medium for data delivery and storage of data is magnetic tape, except for near-real-time systems. As discussed in Section 1.2, the users interviewed want to eliminate these tape libraries and the maintenance costs associated with them. For their applications, the most efficient means would be a direct data transfer from the data base to the user's disk or main memory. The data can be transferred when needed, eliminating the need to make requests in advance and to store copies of data sets. As discussed in Section 3.1.3, other delivery modes may be more appropriate to other user groups (Refs. 8, 10).

2.4 SOFTWARE

Each of the users interviewed has a programming staff, ranging from one to twelve, that develops and/or modifies the application programs that manipulate data for the user's application. The most significant problems concern physical and logical reformatting of data. The physical format of a data

set is the actual arrangement and spacing of bits on the storage device of a particular manufacturer (for example, one cannot mount IBM disks on Univac drives and vice versa). The logical format is the size, type, order, and relationships of the component data items in a data set (for example, binary vs. integer vs. real vs. character). These programmers spend a significant portion of their time reformatting tapes and modifying application programs (Refs. 8, 10). This activity takes valuable manpower away from the actual research analyses.

A few users already have data base management systems and/or extensive software capabilities (Refs. 8, 14). Data users with small budgets, however, may be limited in software capability. As a result, more of the research work has to be analyzed manually. For these users, research capabilities may be significantly improved with the availability of sophisticated software functions (for example, data manipulation techniques). These are functions in which the utility may currently be outweighed by costs; for example, programmer time, processing time, and software expenses (Refs. 8, 10).

SECTION 3

DBMS USER REQUIREMENTS

3.1 DATA SELECTION AND DELIVERY

3.1.1 Interactive Access

The users have all commented about delivery delays that have resulted in the loss of considerable research time. Frequently, data sources and scientific contacts are difficult to find, and documentation about data sets, if available, is sparse (Refs. 8, 10, 14).

The volume of oceanographic data is increasing dramatically, and many users recognize the need for immediate and frequent access to this data base (as defined in Section 2.2). These users want the capability to browse through a data base before making a selection. This addresses a very common problem mentioned by users; the need to enhance their capabilities to locate data that is both available and of good quality. Also, users would find it useful to browse compressed images at reduced resolution (Refs. 5, 8, 9, 10, 12).

It would also be useful to check the availability of other essential data sets for the same time period and region [for example, a research study that requires Seasat Altimeter, Seasat Scatterometer (SASS), and Synthetic Aperture Radar (SAR) data]. Many users rely on the data producer for information; others find it faster to ask for the entire data set than to ask the distributor to be selective (Refs. 8, 10).

Oceanographic data users, particularly frequent data users (as defined in Section 2.1), do not want to incur long data delivery times. A user that accesses data weekly wants to analyze the data as soon as possible, so that preliminary analyses can be completed and used to determine the next data request (i.e., the next day, week, or month). On-line access and delivery would permit the investigator to analyze the data at the time of request, allowing immediate modifications, if necessary, and maximizing the efficiency of the interaction between the user and supplier (Refs. 8, 10).

For users who do not need the capability to browse, such as many commercial users, verbal data requests and changes to the composition of the data sets are sufficient. Although commercial users generally receive data more frequently and uniformly, the data sets vary little, if any, in content, except that near-real-time data is updated every 12 hours (Ref. 11). Conversely, research data requirements are dynamic, and usually each request is different. Exceptions within the commercial sector are special applications in which the data sets must be changed frequently (Refs. 9, 11, 14).

In other disciplines, written or verbal requests may be the most practical for alternate reasons. Interactive access may not be cost effective due to the low frequency of data requests, small size of the user community, or lack of delivery time constraints. Browsing certain types of data may not be productive (for example, SAR photo-negative data) (Refs. 7, 9, 13).

3.1.2 Multi-Key Selection

The interviews revealed that multi-key data selection would be very useful. The users stated that the capability to specify data sets more precisely would reduce editing and significantly enhance the utility of the data. Principal keys for data selection are time, geographic location (space), platform (satellite), sensor (instrument), parameter type, and resolution (temporal and spatial). Other possible keys are data level, principal investigator, and funding agency (Refs. 4, 5, 7, 8, 9, 10).

3.1.3 Multiple Delivery Modes and Products

The diversity of applications and data types demands a variety of output forms: terminal (tabular, graphic), computer-to-computer transmission, and mail (plot, photo negative, magnetic tape). The terminal provides flexible and immediate access to a data base from a remote location. Users can make interactive queries, browse through data sets to determine the suitability of data for their application, and receive products and documentation. Direct computer transmission allows users to transfer data from the host system to the user's disk. This is especially important for frequent data users who require large volumes of data and who want to analyze the data on their own system.

It will still be necessary to mail some types of data. Users lacking plotter capability on their systems should be able to make plots on the host system and have them delivered by mail. It may be more cost effective to mail image data products than to incur the costs of direct transmission, especially if the transmission speed is too slow for effective use. It may not be cost effective to provide on-line facilities to users intending to access the system infrequently; funds spent on terminals and communication lines used once a month could be better spent elsewhere. Other products, such as a SAR photo negative, take time and special equipment to produce and are usually delivered by mail (Refs. 4, 8, 10, 11, 12, 14).

3.2 USER INTERFACE

3.2.1 User-Friendly Command Language

The command language must be simple, concise, and easy to learn and use. For users unfamiliar with the system, menu-guided prompts should assist wherever possible, including a simple help function which should display and explain options. The help function should be invoked voluntarily to avoid frustrating frequent users with verbose prompts. The procedures should be logically ordered, simple, and streamlined; they should not require the user to read operation manuals to do simple functions (Refs. 4, 5, 8, 10, 14).

3.2.2 Flexibility and Compatibility

The DBMS must be flexible and able to easily incorporate new selection keys, commands, functions, parameters, and manipulation capabilities.

The changes should be simple to accomplish, transparent to the user, and have minimal impact on the other components of the data base. The DBMS must be capable of expanding to meet requirements not currently envisioned or cost-effective (for example, changing requirements for a flight project). The system should be compatible with the users' programming languages so that users can invoke pre-established data management functions from their programs (Refs. 4, 7, 8, 10).

3.2.3 Data Independence

The user must be insulated from the data storage media; for example, one should not have to know the precise physical location of the data. Rather, one need only describe the logical parameters (for example, spacecraft, longitude, latitude, time) to access data. The user must be insulated from changes in the data structure and must still see essentially the same outputs, despite internal structural changes (Ref. 14). The purpose is to reduce costly modifications to application programs resulting from changes in the format, mission, or sensor (Refs. 4, 7, 8, 10).

3.2.4 Data Shareability

The DBMS must allow multiple users to simultaneously access the same data, provided their operations do not interfere with other users. This capability is particularly critical with near-real-time data as the demand and value of the data decrease with age. In other applications, users would find it undesirable to wait for other users to complete their data requests, especially if the user community is large and demanding. These problems should be minimized by allowing concurrent access to the data base with minimal delays (Refs. 8, 10, 11, 14).

3.3 DATA MANIPULATION

3.3.1 Conditional Search and Retrieval

Conditional search and retrieval techniques use selection keys (Section 3.1.2) that provide extensive control and flexibility in accessing data. Controlled searches are accomplished by providing limits or specifications for the keys (for example, "list locations where altimeter wave height exceeded 20 feet for September 5, 1978"). As the files in a data base are linked by explicit or implicit relationships, portions of a data base that are relevant to a particular application can be found and retrieved extremely efficiently. The procedure of reading through a magnetic tape library for specific data sets is expensive and time-consuming; computer and operator times are increased for each additional condition (Ref. 1). The users want to reduce costs and processing time. The improvement in data selection capability will allow users to conduct studies in which the processing and manpower costs would be otherwise prohibitive (for example, finding occurrences of oceanographic phenomena in global data sets) (Refs. 7, 8, 9, 10).

Perhaps the most significant impact on research provided by these selection techniques involves the infrequent user with limited resources. Many potential users are prevented from using oceanographic data because they lack the processing facilities, processing time, manpower and/or funds to read and sort the data. Some academic users fit into this category (Ref. 8).

As mentioned in Section 3.1.1, the commercial users receive relatively uniform data sets. In their NRT applications, data manipulation techniques on the host system are not used, as there are severe time constraints on NRT commercial operations. For example, commercial weather forecasting companies must deliver products in a timely manner or lose their contracts to competition. Because the data requests are relatively uniform, search and retrieval techniques would decrease the efficiency of data delivery by adding additional and unnecessary processing steps. Many users have adapted their weather models to the delivery format of the data center(s). Procedures have been developed to put the user in synchronization with the host data source, and thus minimize data delivery time and user-processing (Refs. 11, 14).

For NRT and research applications that are specialized and/or regional, data manipulation techniques may be useful for operations and historical studies. Usually, these applications are short-term or variable; features that make it not cost effective to develop extensive software routines. For example, a 20-year history of severe weather conditions is usually conducted only once for a particular site (Refs. 11, 14).

3.3.2 Reformatting

The ability to efficiently reformat data is important to the oceanographic data users. It is considered the most frequent operation currently being performed by their programmers. They prefer to change the format of the data rather than alter application programs, unless it involves a new mission (e.g., TOPEX) (Ref. 8). In other circumstances, some formats may consume a programmer's time for 3 weeks, including time for inquiries (Ref. 10). This function is required to reduce time spent on adapting programs to new data types and to allow programmers to conduct more study-related programming (e.g., modeling) (Refs. 7, 9, 14).

3.3.3 Quality Control

The quality of the data is a very important and ongoing concern to the users. The users anticipate that the data will have errors; therefore, an overall quality assessment of the data set (for example, from initial investigations) and detailed explanation would be helpful. Users should be able to compare data sets to check quality (for example, compare satellite data to surface data of established reliability) (Refs. 8, 10).

3.3.4 Data Corrections

Users should have the capability to correct specific data points or ranges of points in the data set to be delivered, but without altering the

original. For example, a sensor may have been miscalibrated during part of the mission, or an observer may have taken measurements in the wrong units (i.e., meters - feet) (Refs. 8, 10).

3.3.5 Data Selectivity

The user should be given flexible extraction functions to filter out extraneous data. For example, it may be desirable to remove data points that would have otherwise required tedious geographic coordinate specifications. Other uses consist of removing land contamination, sensor anomalies, and reports from specific vessels (Refs. 8, 9, 10, 11).

3.3.6 Cross-Referencing of Files

Cross-referencing is essential to trace data utilization, development, and relationships between data sets. Because data files can be linked in a data base, it is possible to search for relationships of interest (for example, finding related data files from the same observer or platform). This function can also be applied to literature searches and analytical techniques (Refs. 7, 8, 10, 14).

3.3.7 Compare/Sort/Merge Data Sets

Compare, sort, and merge are functions needed to manipulate data sets into enhanced forms for research applications. A compare function provides the means to evaluate related components of different data sets (for example, values for the same time period and region). Sort provides the capability to order the components of a data set on the basis of field values (for example, order items by increasing latitude); the fields can be selected through a sort key (for example, wind speed, time, etc.). Merging data sets is useful for creating continuous, consistent, homogeneous data sets from separate but related data sets (for example, to display wind data on an image data background). Other capabilities consist of overlaying data sets graphically, averaging, interpolating, and generating gridded data sets (Ref. 2).

Currently, these functions are accomplished using inefficient computer procedures that compare each element of both data sets. Sometimes, these procedures are performed manually. For many users, the manpower and processing costs outweigh the utility of the end product. The strawman in Appendix B provides a list of common procedures to merge data sets (Refs. 8, 10).

The organization of a data base linked by relationships is more efficient, provided the keys (see Section 3.1.2) have been selected correctly by the system designers. This organization would significantly facilitate research work by providing the data sets in forms that can be readily analyzed by scientists (Refs. 4, 8, 9, 10, 14).

3.3.8 Graphics

Visualizing data analyses through graphics is easier to understand and use than other media. Outputs can be produced on graphics terminals, image terminals, or multi-colored plotters. Graphics should not be considered an alternative to "numbers" but rather as a supplement. These visual tools can help significantly with initial analyses of data (e.g., trying to identify subtle phenomena in a global data set) (Refs. 4, 8, 9, 10, 14).

3.4 DOCUMENTATION - DATA DICTIONARY

At present, users have to rely on catalogs or data center contacts for information about data sets. The appropriate data source or scientific contact may take time to locate. Catalogs must be mailed and frequently have insufficient information. These delays are no longer acceptable because of increasing data volumes, processing, and user data requirements.

As discussed in Section 3.1.1, the users need an on-line data dictionary or catalog for browsing and literature/data searches through the data base. Documentation is provided in machine-processable form pertaining to each data item in the data base and to each program that accesses the data base. The data dictionary itself is a data base; interactive queries are made to extract the desired information and documentation. Many users do not have the time to read or search through lengthy catalogs. Efficient dissemination of useful information is greatly needed and required by the oceanographic community. Information on software development, problems with data sets, and complementary data sets can save an investigator considerable research and programming time.

Some of the information categories needed by the user community consists of:

- Source (for example, Seasat altimeter).
- Coverage (geography and time).
- Resolution.
- Frequency (data collection).
- Quality assessment.
- Related data sets (for example, surface observation data sets for sensor data files).
- Algorithms (sensor data processing).
- Description (for example, technical description of sensor).
- Availability (of particular data sets).
- Processing time and cost (for user data requests).

- Status (of a data set, extent of processing).
- Notes on missing or poor data sets (comments section).
- Project software documentation.
- Tape or disk numbers and associated coverage.
- User application programs (who, what function, availability).
- Scientific contact.
- References.

(Refs. 4, 5, 7, 8, 9, 10).

3.5 OTHER SYSTEM CAPABILITIES (SECURITY, PRIVATE FILE, PRODUCT ARCHIVE, REPORT WRITER, APPLICATION PROGRAM)

Users need system services to supplement data extraction functions and to increase the utility and capability of the data network. The user interviews revealed that security, private file, product archive, application program, and report writer capabilities would be useful to the oceanographic community. In particular, users that lack extensive processing capability would realize the most benefits.

The most essential services identified by the users are security and private file capability. The system security must protect the data base and limit user functions and updating privileges. Private file capability should be incorporated to enable users to store limited private data sets on the system. This capability should be provided for a system-specified period of time. One must be able to limit access to private data files to specified users.

Product archive capability is essential to store subsets of data files generated by users. This service would allow a user to store an unfinished data product and permit further work (i.e., filtering out extraneous data, comparisons, merging, etc.) at a later time.

A direct interface between the data base and user application program would significantly increase the efficiency of user processing by eliminating manual data transfer operations. Users should be able to execute limited application programs on the system. This may be impractical, however, if the demands on the system are significant. A more efficient system capability would involve remote access to the data base from the user's computer; the data extraction commands would be built in to the user's application program (see Section 3.6.5).

A report writer would be useful for outputting data files or computations in a report format. This could include a graphics feature for incorporating graphs or plots. These reports would be used in publications and presentations (Refs. 4, 8, 10, 14).

3.6 PERFORMANCE

3.6.1 Response Time

As stated in Section 3.1.1, oceanographic users need interactive access to the data base. Many users will browse through the data base, retrieve data sets, and display the data on their terminal or transfer it directly to their computer. It will be detrimental and counterproductive if users have to spend excessive time waiting for system responses. The users want to select and retrieve data, including image data, quickly and easily. This capability will allow them to spend most of their time on more critical activities, such as analyzing the data. To accomplish this, all user operations should be processed efficiently to minimize response time. Concurrent use of the system should not cause any significant delays in the response time. Because the potential user community is sizable and data volumes are rapidly increasing, it is critical that the system have the capacity to handle these demands. Fast response times are very important to many oceanographic data users, especially to those that foresee data access and analysis on a daily basis (Refs. 4, 5, 7, 8, 10, 11, 14).

Besides data selection and retrieval, some operations require significant processing which may cause a delay in response time. Examples of this are some data manipulation techniques and application programs. The users should be able to perform these operations in a background or overnight mode. This ability would enable a user to perform other operations on a terminal while waiting or to perform these slow operations when not connected to the system. These capabilities would make it unnecessary to sit idly at a terminal while an operation is being completed.

3.6.2 Processing Time

Data requests must be processed efficiently to effect timely data delivery. As discussed in Section 3.6.1, users do not want to incur slow response times because of inefficient processing. Also, users want to minimize processing time fees, if applicable (Refs, 4, 7, 8, 10, 14).

In addition, the data itself must be processed efficiently. In most cases, data is useful only if incorporated into the system within a reasonable time after collection and initial processing. The degree of timeliness varies according to the application, but delays have caused severe difficulties for even long-term research projects (Ref. 10). The DBMS must have the capability to process and store the large data volumes anticipated in future systems (Refs. 8, 10, 14).

3.6.3 Availability

With the requirements presented in 3.6.1 and 3.6.2, the DBMS should perform all of the necessary system and user functions without any degradation of system availability. The effects from a large user community accessing large data volumes must not be ignored, especially when considering data

volumes from future satellite systems. Under-designing system software for projected data requirements will significantly lessen capacity by limiting user access (Refs. 8, 10, 11, 14).

3.6.4 Data Integrity

The DBMS must maintain the data precision or accuracy. This should include monitoring for unreasonable values, inconsistencies to known relationships, data gaps, and incorrect data structures. Any decimation or regridding should be transparent, and either reversible or access made available to the original data set (Ref. 10). Related requirements include error prompts and messages in the user interface (Section 3.2.1), data set comparisons (Section 3.3.3), and documentation (Section 3.4). Data sets will be useful only if they are of sufficiently high quality (Refs. 4, 7, 8, 10, 14).

3.6.5 Distributed Networks

For the oceanographic community, it would be desirable to tie all oceanographic data systems together. Data base management systems could be interconnected, allowing a user to receive data from multiple sources at several nodes in a distributed data base network. The user would not be required to know the computer system(s) on which the requested data resided. A distributed processing network, alternatively, would allow a user's application program to access a remote data base using its resident DBMS and then process it using application program(s) that may reside on a different processor. Both types of networks would facilitate oceanographic data studies. A user could access data from more than one data center, use multi-center catalogs, and run application programs that access remote data bases. Large application programs could access the data base without burdening the DBMS or data center processor. Conceivably, the user could retrieve and use any available data with minimal effort; data collection procedures would be greatly simplified (Refs. 7, 8, 10).

3.6.6 Near-Real-Time (NRT)

The DBMS should have the capability to store and deliver NRT data in a timely manner. These applications also require a higher degree of availability and response time performance. As stated in Section 3.3.1, most data manipulation functions are of limited use on a NRT data distribution system, especially time-consuming functions like overlaying and merging. A DBMS, however, can be used for specialized applications studying phenomena or specific geographic regions. The purpose is to design a generalized DBMS that is flexible and compatible with future systems and that can be incorporated into NASA systems, including near-real-time systems. These requirements are directed toward the development of DBMS software; hardware and operating systems, with the exception of DBMS compatibility requirements, are outside the scope of this study.

Oceanographic data systems are being developed to handle the real-time data streams from satellite systems. As a result of these developments, many scientists foresee more NRT applications. Many of the techniques and programs that they develop will be used in real-time; these must be tested in NRT operating environments (Refs. 8, 11, 14).

SECTION 4

USER INTERVIEWS/SUMMARIES

James Marsh
Vincent Grano

Goddard Space Flight Center
November 1980

OVERVIEW

User activities

Ocean circulation, topography, and geodetic studies utilizing Seasat altimeter, GEOS, surface buoy, bathymetric, and surface temperature data. Altimeter data is being used to study Gulf Stream system. Studies are published in NASA documents.

Data suppliers

Centers: JPL, Wallops, NOAA, Woods Hole, Scripps, and others.
Form of data requests: By letter or telephone
Selection criteria: Time, location, and sensor.
Additional selections: Level 0-1-2, grid size (i.e., 63 x 63)
Present form of outputs: Magnetic tape
Format: Since some programs were developed five years ago, it is easier to reformat data sets to the in-house format. However, format and program changes are required when the existing format is not comprehensive enough (for example--TOPEX has new parameters).

EQUIPMENT

Computers: IBM 360/95 360/91 (these are to be replaced by a vector-processing machine).
Data Storage: Tape, disk
Access capability: Batch
Desired media for data deliver: Terminal, computer
Comments: Users need on-line capability to collect oceanographic data from data centers and thereby cut turnover time. Future systems should store massive amounts of data on-line and have the power to process large volumes of data.

SOFTWARE/APPLICATION PROGRAMS

Access language: FORTRAN

Application programs: Process raw data to extract certain quantities of interest in order to study circulation and kinetic energy. Both standard and special programs are utilized.

Modifications to application programs: Application programs are modified to enhance their capability (for example, increase capacity from 10 to 100 passes). New application programs start with a skeletal design and are developed into more powerful and efficient programs as the need arises. Format changes to these programs are infrequent as the users prefer to change the format of incoming data.

Manpower: Four to five programmers are utilized on a half-time basis. The computing budget ranges between $100-200K per year. It was estimated that the programmers spend approximately 30% of their time performing tape operations (for example, loading, copying, reading), manipulating data sets (for example, reformatting, extracting, sorting), and working with catalogs.

Software: Operating system - IBM, MVT (multiple variable number of tasks)
IBM data manipulation software - SORT function
Plotting routines - CALCOMP

DATA MANIPULATION

Conditional search and retrieval: Useful for collecting satellite passes in certain areas of particular times. Concerning extraction of useful data from data sets, the users prefer to perform these operations on their facility and under their control.

Reformat: Very important. This would save considerable amounts of processing time and cost.

Data corrections: Prefer to control themselves. However, they would like access to and use of the algorithms used to process raw data (for example--as with Seasat).

Compare/Sort/Merge: These tools would be extremely useful considering the different types of data used in their studies: altimeter, bathymetric, geoid, mean sea surface, pressure, wind, wave, field measurement (dynamic topography), current, sea surface temperature, GEOS image, and storm data.

Overlay: The capability to overlay image data would be very useful, especially storm data. These operations may be performed daily.

Comments: The users stated that all the manipulation tools would be useful to their research and are necessary to visualize and examine the data.

INFORMATION

The users stated that all of the information types listed in the outline could be helpful with the following additions and notations:

(1) Scientist to contact and his/her telephone number.
(2) Algorithms with references.
(3) Usage of data: Who has used the data.
(4) User problems: Ability to determine all users of the data set that have experienced a particular type of problem with the data.
(5) User application programs: Information on user-developed application software may prevent redundant development or duplication of effort.

SERVICES

All of the services listed in the outline would be useful including archive, report generation, edit, user data entry, application programming. Archive capability would be useful for making data sets available to other scientists working on the same project. The report generator would be useful for preparing charts for publication. Security and private file capability would be especially important for controlling access to private data sets and application programs.

PERFORMANCE

Timeliness, fast response times, and high availability are essential; they are the reasons for developing more powerful systems. An interactive system with on-line data delivery would be a tremendous improvement over present systems. The ability to quickly determine what is available and to scan data sets would save considerable time.

ESTIMATED USE

Frequency: Probably daily; difficult to specify since dependent on the study.

Utility for NRT data: Even though charter is primarily research, users foresee real-time applications (for example, NOSS). In addition, some of the techniques and programs that they develop will be used in near-real-time.

Data sets to be utilized in the future: Future studies will use data from TOPEX, NOSS, Seasat, surface buoy data (density, salinity), SAR (optically reduced form), etc.

Michael Parke
Jet Propulsion Laboratory
December 3, 1980

OVERVIEW

User activities

Analyzing tides and ocean currents. This consists of studying spin-up of eddies off Somalia (Indian Ocean), tides on the Patagonian shelf off Argentina (and globally), and the interaction of winds and currents in the Drake Passage (between South America and Antartica). Data requirements for these studies consist of altimeter, SASS wind fields, bottom pressure, atmospheric pressure, wind, current meter, and other surface observation data. A former activity involved Seasat altimeter processing.

Data suppliers

JPL (altimeter and SASS) and personal contacts (see comments below).
Form of data requests: By letter or telephone.
Selection criteria: Time, location, sensor, and parameter.
Present form of outputs from data sources: Magnetic tape, tables, and plots.

Comments: Data selection depends on availability, interest in region, and quality of the data. The user may be informed of good quality data in an applicable region (for example, off Somalia or Chile) through personal contacts. These personal contacts may in turn have learned of or received these data sets through their own contacts.

EQUIPMENT

Computer: Prime 550, CDC 7600.
Data storage: Tape, disk (approximately 50M bytes at disposal).
Access capability: On-line.
Desired media for data delivery: Terminal, computer-to-computer, magnetic tape.

Comments: The most significant problem is the delay in data delivery; a study was terminated as a result of this problem. Although the user's frequency of access may vary, he needs the interactive capability to assess the quality and suitability of the data sets (i.e., make sure numbers are correct and capable of supporting study), to perform initial analyses (i.e., plots and listings), and to edit data sets to filter out undesirable data.

SOFTWARE/APPLICATION PROGRAMS

Application program language: FORTRAN.

Models and analytical techniques: Tide and ocean current models tailored for each project. Global modeling routines are split into many subprograms.

Modifications to application programs: Analysis routines often last for one project. Depending on the new project, old routines may be usable with small changes; in others, the core subroutines will be reused. The form and sequence of data inputs (logical format) or headers have to be changed on modeling programs at these times. Programs continually evolve. General utility routines may be applicable for other projects and thus have a longer life than analysis routines (i.e., a plot routine vs. a tide model).

Effects of changes in the data base, mission, format: Expensive because much programmer time is required for each change. However, this depends on the scope of the project. The application programs may be tailored to the mission (i.e., Seasat) because the project is limited to a few data sources. As mentioned before, analysis routines usually last for one project; a new project may require extensive modifications to application programs as a result of a new mission (i.e., TOPEX) or a change in the focus of an on-going study.

As the data sources vary, physical format modifications are required on many data sets before being used in application programs (i.e., interpreting, sequencing, offsets, binary integer conversions). Some formats are very difficult to interpret and consume a few weeks of programmer time (i.e., one format took a programmer three weeks to decipher, including time for inquiries to the data source).

Manpower: One half-time programmer. Computing budget from project funds.

DATA MANIPULATION

Conditional search and retrieval: Very useful for research applications. This could be used to locate interesting data (i.e., data during upwelling) or to find regions and times in which necessary data sets are available (i.e., Seasat altimeter, Seasat SASS, current meter data).

Reformatting: The ability to easily reformat data sets, both physically and logically, would be very useful. This would lessen problems in deciphering formats and may make it easier to use application programs for new projects.

Quality control: Good quality data is essential for user's application. The capability to assess the quality of available data sets quickly would make data selection much easier; at present, the user must rely on the data source for this information.

Removal of extraneous data: Useful since user could refine data set to essential components (i.e., filter out unnecessary bits or bad data).

Cross-referencing of files: Useful for locating related data sets (i.e., finding surface data collected simultaneously with sensor data or finding other data collected by same observer).

Compare/sort/merge: Very useful to be able to perform these operations but not necessarily on-line. Comparing data sets would be useful for scientific applications and quality assessment. Concerning merging techniques,

the user stated it is important to know precisely how data files are merged. The capability to select grid sizes (i.e., 63 x 63) would also be useful.

Graphics: Simple graphics are very important for scientific applications; however, fancy plots (i.e., overlay) would be useful. Plots can be sent by mail.

INFORMATION

All of the information types listed (see Appendix A) would be useful, especially documentation on the algorithms and processing. The documentation should be on-line and easily accessed. The information should help a user to select and use the data.

SERVICES

All of the services would be useful: archive, report generation, edit, user data entry, and application programming. Security and private file capability would be especially important (i.e., controlling access to private data sets and application programs).

PERFORMANCE

The DBMS should be flexible to changes in requirements and easy to use and learn. The most significant problem is waiting to receive data. The system should be efficient: fast response times, high availability, and sufficient processing capability. In addition, a distributed data base network would be very important for facilitating data selection and delivery from different sources.

ESTIMATED USE

Frequency: Difficult to specify since dependent on the project.

Utility for near-real-time data: A few applications. For time-dependent activities, it would be useful. That is, it would be costly to send out a ship at random to examine a variable upwelling. Infrared pictures can be used to identify start of upwelling and hence, when and where to take measurements.

Warren Yogi
Rodger Langland

Ocean Data Systems, Inc.
Monterey, CA

OVERVIEW

User activities

Specialized weather forecasting for government, scientific, and commercial groups; work is performed on contract. Participant in Seasat ASVT experiments and user of joint NASA (JPL) - U.S. Navy (FNOC) Satellite Data Distribution System (SDDS).

Operational mode - real-time weather forecasting and archived data analyses. Research and development mode - studies for JPL and DOE.

User's data base

Maintains surface data for one-month period; however, some experimental data kept for two months. User has easy access to FNOC data base, and most applications utilize current data only.

Real-time vs. archived data

Principal operations are on-line and real-time. Products must be delivered in short time frame to customers. No time to study data. Uniform data set delivered every 12 hours. Archived data is used only if a contract requires it to perform analysis.

Data suppliers

FNOC: SOWM, wind, wave, SST, 500 Mb pressure
Nimbus 7 SST data.
FNOC archived data base.
WMO-GTS (Global Telecommunication Service)
RAWARC (Continental U.S. - radar reports, tornado)
AG NET (Agricultural network)
NWS (National Weather Service)
West Coast Marine Circuits (to be connected)

Form of data requests: For applicable sources, requests are made by telephone and/or mail (for example, SDDS). Data streams are uniform, and sufficient notice can usually be given for changes.

Selection criteria: In general, data has been preselected; changes are infrequent, usually when a contract is initiated or terminated. However, for specialized applications, it may be desirable to interactively select data by keys: source of data, location, time period, and parameter.

Present selection criteria: By region, time, and parameter.

Present form of outputs:
Real-time operations: computer-to-computer transmission. Retransmit

data to main system in Minneapolis where analysis is performed. Tabular and graphic displays sometimes used for special applications.

Archived data analyses: magnetic tape

Community served: Entire marine community - contract dependent.

Comments: Features are desirable if they lessen user's processing time without sacrificing overall timeliness of data. User prefers to perform analyses on data himself, except for refining data sets.

EQUIPMENT

Computers: two CDC Cyber 1820.

System A - Message switch system (receives inputs and stores data).

System B - Transmits data to Minneapolis and receives forecast models. Plot and batch capability.

Peripherals - Disk and magnetic tape drives, printers, terminals, plotters.

Minneapolis system:

CDC Cyber 203 - Array Processing Machine.

CDC Cyber 175 - Front-end to 203 (user interface-dial-in parts).

Data storage: Disk, magnetic.

Access capability: On-line and batch.

Media for data delivery:

Real-time: Utilize computer-to-computer transmission of data wherever possible. On-line access.

Archived data: On-line desirable, presently magnetic tape.

Interface requirements: Computer protocol formats must be CDC compatible.

Comments: If user can rely upon a source, changes in user's software will be justified.

SOFTWARE/APPLICATION PROGRAMS

Access language: FORTRAN.

Operating System: SCOPE 3.4, NOS.
MSOS on systems A & B.

Standard vs. special programs: Both are utilized.

Standard packages are the global forecast models.

Specialized application programs are designed for specific contracts or for analyzing specific storm systems.

Modifications to application programs: Modifications are infrequent unless required by contract or procedure. Their data sources rarely change formats (see next section).

Effects of changes in the data base, mission, format: Extremely expensive as models are quite extensive. These changes are rare.

Mission data is limited to studies. The changes in the Seasat data did not affect the analysis time significantly. The user is in the process of changing formats; one data source wants to change to the spot format. Large software changes will be required.

Manpower: Twelve full-time programmers. The programming staff spends nearly all of its time developing new application programs.

Computing Budget: Approximately $500K per year. Provided by contracts.

Software Packages: Designed own DBMS for system: MSS. Customers can make interactive queries to ODSI data base for station, parameter, and time (for example, SFO winds OZ).

Comments: The user's primary functions utilize global data sets to perform global weather analyses. The ADBMS would be more useful for the specialized application programs involving specific regions or parameters. The incentive for an ADBMS would be to reduce processing time and to facilitate specialized and archived data analyses.

DATA MANIPULATION

Conditional search and retrieval: Useful only for studies and some special application programs. Not useful for global data sets. To be used operationally, timeliness is essential. However, user processing time for data is short in relation to transmission time; moving data is the most time-consuming.

Reformat: Not critical as ODSI stated it can receive almost any format.

Quality control: Essential as do not have time to check data, but there is no time to examine or manipulate the data sets anyway. Suggest use of flags on data and/or bulletin file. The flags can consist of data quality bits. The bulletin file can provide figures giving the number of raw observations received against the number of bad observations.

Removal of extraneous data: Performed by user's system for some applications. Operationally, all data is utilized quantitatively.

Cross-referencing of files: Helpful for special application programs and studies.

Compare/sort/merge: Useful only for studies as do not have time with other applications, except if it saves processing time and delivers products within acceptable time limits.

Overlay: Performed for studies.

INFORMATION

All of the information types listed would be useful, expecially quality assessment and usage of data.

PERFORMANCE

Timeliness, fast response times, and high availability are essential. Real-time constraints. Ability to move large data sets quickly and accurately.

ESTIMATED USE

Frequency: Every 12 hours (raw weather observations)

Utility for near-real-time data: Forecast models (operations)

Data sets to be utilized in the future: Nimbus 7 (SST), NOSS, TOPEX.

SECTION 5

REFERENCES

1. Apenyo, Kofi, Survey of Representative Database Systems, Draft, Jet Propulsion Laboratory, Pasadena, California, November 1, 1980.

2. Cardenas, Alfonso, Data Base Management Systems, Allyn and Bacon, Boston, Massachusetts, 1979.

3. Fujimoto, Brad, and Renfrow, Tom, Seasat Satellite Data Distribution System: Interim Evaluation Report, Internal Document No. 622-219 Rev. A, Jet Propulsion Laboratory, Pasadena, California, April 15, 1980.

4. Gary, J. Patrick, et al., NASA End-To-End Data System (NEEDS) Data Base Management System Functional Requirements, Information Extraction Division, Goddard Space Flight Center, Greenbelt, Maryland, March 1980.

5. Lame, David B., "Ocean Data Utilization System Study Progress Report #2," Internal Memo, Jet Propulsion Laboratory, Pasadena, California, November 26, 1980.

6. Lohman, Guy, Functional Requirements for an Applications Data Base Management System, Draft, Jet Propulsion Laboratory, Pasadena, California, November 15, 1980.

7. Lohman, Guy M., and Renfrow, Thomas J., A Proposed Concept for a Crustal Dynamics Information Management Network, Publication 79-111, Jet Propulsion Laboratory, Pasadena, California, January 15, 1980.

8. Marsh, James, and Grano, Vincent, Goddard Space Flight Center, Personal communication, November 19, 1980.

9. OAO Corporation, User Requirements for NASA Climate Data Base Management Systems, prepared for Goddard Space Flight Center, Greenbelt, Maryland, under contract NAS5-23436 Mod. 27, OAO Corporation, Beltsville, Maryland, October 1, 1979.

10. Parke, Michael, Jet Propulsion Laboratory, personal communication, December 3, 1980.

11. Seasat Satellite Data Distribution System, Jet Propulsion Laboratory and Fleet Numerical Oceanography Center, personal experience, August 1979 to September 1980.

12. User Requirements Support Package for the OSTA Data Systems Planning Workshop: Geodynamics, conference held at Wallops Island, Virginia, October 1979, prepared by Ecosystems International, Inc.

13. User Requirements Support Package for the OSTA Data Systems Planning Workshop: Non-Renewable Resources, conference held at Wallops Island, Virginia, October 1979, prepared by Ecosystems International, Inc.

14. Yogi, Warren, and Langland, Rodger, Ocean Data Systems, Inc., personal communication.

APPENDIX A

OUTLINE FOR USER INTERVIEWS

A. OVERVIEW

User activities and products (operations)
Types of oceanographic data utilized
- surface observations, numerical models
- low-rate satellite data
- high-rate satellite data

Specific data sets utilized
Other data utilized (user's database)
Real time vs. archived data
Data suppliers (location, media, format, preprocessing, accounting)
- form of data requests (terminal, memo, verbal)
- preferred data selection criteria (keys) (source-mission, time, location, sensor, measurement, parameter, units, resolution).
- form of outputs
  - tabular, graphic, photo-negative (graphics capability)
  - essential components of data set (instrument dependent)

Community served (results issued, data forwarded)

Does the user currently obtain sufficient quantity and quality of data for his application?

B. RESOURCES - COMPUTER FACILITIES

Equipment
Data storage (tape, disk, etc.)
Access capability (on-line, batch)
- desired media for data delivery (mail, terminal display/ printout, tape)

Interface requirements

C. SOFTWARE/APPLICATION PROGRAMS

User's access language
Models and analytical techniques
Standard vs. special programs
Type and frequency of modifications to application programs
Effects of changes in the database, mission, format
Manpower
- how many programmers?
- what are they doing? (reprogramming?)
- turnover

Computing budget
- how much?
- where from?
- how allocated?

Software packages used by user's system (GFMS, etc.)
Future software developments
- where is user going in terms of software expenditures?

D. **DATA MANIPULATION**

Conditional search and retrieval
Reformat (preferred format?)
Quality control, inconsistencies in data
Removal of extraneous data (type?)
Cross-referencing of files
Data corrections
Comparing/sorting/merging data sets

- Earth grid selection

Manual operations (overlaying, graphics)
What kinds of operations are frequently performed?
How is the data being used?
Why are these operations necessary?

E. **INFORMATION - DOCUMENTATION**

Information sources
Useful information about data files

- source, coverage, resolution, frequency, quality assessment, related data sets, algorithms, description, availability, access, scientific contact, references
- processing time and cost to the user
- status
- notes on missing or poor data sets
- access to level 0 or level 1 data (SASS)
- project software documentation (chickens)
- tape numbers and associated coverage
- user application programs (what, where, availability)

F. **SERVICES DESIRED**

Archive, report generation, edit, user data entry, comprehensive computer programming
Private file capability, security

- ability to make data available to selected groups

G. **SYSTEM PERFORMANCE**

Timeliness (if applicable)
Desired response time (to be feasible for applications)
Availability

H. **ESTIMATED USE OF ADBMS**

Typical work load/data volume to satisfy future needs

- anticipated frequency (requests)
- extent of use
- quantity of outputs

Utility for real-time or near real-time data
Data sets to be utilized in the future
Demonstration material

APPENDIX B

STRAWMAN USER REQUIREMENTS FOR OCEANOGRAPHIC DATA MANAGEMENT SYSTEMS: CURRENT SCENARIO

Sample Activity	Examining wave-current interactions and shallow-water wave refraction.		
Procedure	Seasat data to be integrated into a surface measurement data base. This will be used for verification and comparison to numerical wave current interaction and shallow-water wave model studies.		
Data category	High-rate Satellite	Low-rate Satellite	Surface observations Numerical model.
Example data sets:	SAR	Altimeter	Wave height (H1/3), Wind speed and direction Bathymetric data.
Data source	JPL SAR Processing Lab	Seasat Project Data Processing System.	Fleet Numerical Oceanography Center.
Data medium	Photo negative	Magnetic tape	Magnetic tape.
Orientation, format	Geographic satellite pass, asynoptic	Chronological satellite pass,	Chronological grid, synoptic.
Footprint	100 km.	2.4 - 12 km	40-381 km/grid.
Swath position, grid size	20° off nadir	Nadir viewing	63x63 grid.
Delivery time	1-2 months	1-2 months	1-2 months
Delivery mechanism	Mail	Mail	Mail
Programming languages	-	Non-transparent	Non-transparent
Application programs	Dependent	Dependent	Dependent

Query capability: User can only select data.

Satellite data: The desired longitude, latitude, and time period are specified and translated to satellite pass numbers and time.

Surface data: The user selects the desired product(s) along with the desired longitude, latitude, synoptic specification, and time period. The request is translated to a corresponding zone on their conventional grid.

CURRENT PRELIMINARY PROCESSING BY USER

(1) Read magnetic tapes on compatible equipment.

(2) Remove extraneous data and ambiguities.
A user is usually given data in blocks that frequently include data outside the region of interest. In addition, some data sets have bad data.

(3) Reformat data sets.
Frequently, it is necessary to restructure the data into compatible forms for comparison and merging with other data sets.

(4) Sort/merge individual data sets
Order the data geographically into a common grid pattern (geographic proximity formatting).

(5) Correct data
Time corrections may be necesary when data sets involve different collection methods. This is frequently encountered with satellite data that is asynoptic and surface observations that are synoptic.

(6) Merge different data sets
Geographic and time parameters must be compared for each value. Multiple values for the same grid point have to be averaged.

(7) Plot data on map for analysis.

APPENDIX C

STRAWMAN USER REQUIREMENTS FOR OCEANOGRAPHIC DATA MANAGEMENT SYSTEMS: POTENTIAL FUTURE SCENARIO

Data category	High-rate satellite data	Low-rate satellite data	Surface observations numerical model
Example data sets	SAR	Altimeter	Wave height (H1/3), wind speed and direction. Bathymetric data
Data source		DATA NETWORK	
Data medium	OFT plot[1]	On-line terminal[2]	On-line terminal[2].
Orientation	Geographic	Geographic or Chronological	Geographic or Chronological.
Footprint	100km	2.4-12 km	40-381 km/grid
Grid size	Variable[3]	Variable[3]	Variable[3]
Delivery time	1 week[1]	Immediate[4]	Immediate[4]
Delivery mechanism	Mail[1]	Telephone line	Telephone line
Programming languages	-	Transparent	Transparent
Application programs	Independent	Independent	Independent

Query capability: User can select and manipulate data records (delete, insert, change, compare) based upon multiple keys (time, latitude, longitude, sensor, geophysical parameter)

Note 1: An Optical Fourier Transform is one example of higher-level SAR processing. It is possible to receive an OFT in plot form, provided a digital plotter is available at one node of the network. In addition, digitized SAR data can be transmitted on-line. The delivery time and mechanism refer specifically to the use of a plotter.

Note 2: The choice of an on-line terminal is not absolute but convenient. A user can choose to receive data in tape or printout forms. Alternatively, data can be transferred utilizing computer-to-computer (host-to-user) communications.

Note 3: The grid size is specified by the user but must be within reasonable bounds. This capability makes it possible for the user to format different data sets in the same grid pattern.

Note 4: Immediate delivery is possible when utilizing an on-line system. This interactive mode makes it possible to search, request, and receive data in a short time frame.

POTENTIAL PRELIMINARY PROCESSING BY USER

ON DATA NETWORK:

For example, to plot geographic points in which altimeter (ALT) data is within ten percent of surface measurements (H1/3); place inconsistencies between sensor and surface data (greater than 10% difference) in a separate file (B).

QUERY:

(1) MERGE 63x63 ALT, H1/3
(2) REGION 80W30N, 80W20N, 97W20N, 97W30N
(3) TIME 7-1-78, 8-1-78
(4) MATCH ONLY REGION, TIME
(5) ALT >= 0
(6) | ALT - H1/3 | < (0.1 x H1/3) OR ELSE STORE B ALT, H1/3, LONGITUDE, LATITUDE, TIME
(7) PLOT X10, Y10
(8) OUTPUT B, TERMINAL

EXPLANATION OF HYPOTHETICAL QUERY COMMANDS:

(1) Retrieve these data types, select in accordance with specifications below, and merge these data sets for comparison using 63 x 63 grid
(2) Region selection (longitude-latitude)
(3) Time period (July 1 to August 1, 1978)
(4) Only select sensor and surface data in which the region (grid point) and time parameters are the same
(5) Only select reasonable sensor values (i.e., remove land contamination)
(6) Only select points with less than a ten percent difference; place points that don't meet this constraint in file B. File B will contain the data values along with their corresponding longitude, latitude, and time parameters
(7) Plot the points that meet the specifications and constraints in steps 1-6. User may specify plot function parameters (i.e., borders, colors, etc.)
(8) Send file B to user terminal.

USER FACILITY:

- Plot of matched data values is mailed to user
- Table of inconsistent data is received on user's terminal
- Analyze data

EXAMPLE OF DATA DICTIONARY

(reprinted from J. Patrick Gary, et al., NEEDS DBMS Functional Requirements, Goddard Space Flight Center, Greenbelt, Maryland, March 10, 1980.)

MICROWAVE RADIANCE
(As of December 1979)

a. Type

1. Data Set Name
Microwave radiance

2. Parameter/Measurement
Multispectral passive microwave images

3. Source
Nimbus 7 Scanning Multichannel Microwave Radiometer (SMMR).

4. Level
Level I

5. Units
Units, Kelvin

b. When

1. Temporal Coverage - October 1978 to present.

2. Temporal Resolution and Frequency - Every other day, starting November 16, 1978, global coverage every 6 days.

c. Where

1. Spatial Coverage - Global (to latitudes 84° N&S).

2. Spatial Resolution - Footprint is eliptical, the major/minor axis ratio varies with frequency from 151/97 (km) at 6.6 GHz to 27/18 (km) at 37.0 GHz.

d. Instrument Description

1. Satellite - Nimbus 7, a multi-experiment observatory at a nominal 955 km altitude in a sun-synchronous, high-noon orbit.

2. Instrument - The instrument scans to either side of the orbital track, with a conical half angle of 42°, with an angle of incidence on the surface of the earth constant at 50.3°. The scan is sinusoidal and overlap coverage is provided at all wavelengths. Details are given in the Nimbus 7 Data User's Bulletin. A similar instrument flew on Seasat A. Both instruments deliver orthogonally polarized antenna temperature data at five microwave frequencies.

3. Launch Date, Duration - Launched October 24, 1978. Operating as of December 1979.

4. Sensor Scientist

   Dr. Per Gloersen, Code 913
   Goddard Space Flight Center
   Greenbelt, MD 20771

5. Mission Objectives - The scientific objectives of the SMMR experiment are:

   a. Extract geophysical parameters from multispectral microwave radiances.

   b. Verify the extraction algorithms.

   c. Utilize the extracted parameters in climate modeling and assessment.

   d. Support ongoing and new operational maritime uses (Fleet Weather Facility-USN/FWF, Fleet Numerical Weather Central - FNWC).

   e. Identify new observables.

These goals will be achieved through spatial and temporal determinations of such geophysical parameters as:

   a. Integrated water vapor, total liquid water and (possible) rain rates.

   b. Ocean surface temperatures.

   c. Ocean surface roughness.

   d. Wind speeds determined from ocean wave radiances.

   e. Sea ice concentration.

   f. Sea ice concentration changes.

   g. Hurricane and similar adverse weather energy balances computed using experimental data.

6. Principles of Operation - The instrument employs a novel antenna system in which a 42 degrees offset parabolic reflector illuminates a single feedhorn covering the entire range of operating wavelengths and which provides coaxial antenna beams for all 10 channels. The retrieval of geophysical parameters from one of more members of the sets of antenna temperatures is based on the variation of microwave emission reflectivity and absorption with frequency and polarization. (See f).

7. Measurement Geometry - (TBD)

8. Data Recording - Analog data from the radiometer channels are fed to the A-D converters, the converter outputs are commutated in sequence into the shift registers, and the outputs then fed into the spacécraft data stream. Data sampling rates vary between channels, with those channels having smaller IFOV's being sampled more often. The SMMR output data rate is 2 kb/sec.

e. Related Data Sets

SMMR data products are of two kinds: tapes and map displays. The principal product at the sensor level (Level 1) is the Antenna Temperature Tape (TAT) (Section h). A tape archived but not routinely delivered to users is the User Formatted Output Tape-SMMR (UFO-S) which contains the primary sensor data (not calibrated) and the related SMMR telemetry. A third tape is denoted as CELL-ALL. In it, individual fields of view have been composited to three standard cell size in preparation for retrieving geophysical parameters. Map displays are not prepared for antenna temperatures except for the average 37-GHz brightness temperatures.

f. Processing Algorithm

The functional form of the equation which relates Antenna Brightness Temperature T_B to telemetry counts (normalized) is:

$$Tb = A + BN + C\ (t_h - t_{ho}) + D\ (t_h - t_{ho})\ N + (1 - \alpha)\ (t_I - t_{I0})$$

Where:

T_b = Brightness Temperature

A, B, C, D, α = Constants to be Determined

N = Normalized Count

t_h = Hot Load Temperature

t_{ho} = Reference Hot Load Temperature

t_I = Instrument Temperature (e.g., Radome or Antenna)

t_{IO} = Reference Instrument Temperature (e.g., Radome or Antenna)

N is defined as:

$$N = \frac{C_A - C_H}{C_C - C_H}$$

where:

C_A = Digital counts from the radiometer, when viewing the earth.

C_C = Digital counts from the radiometer, when viewing the space horn.

C_H = Digital counts from the radiometer, when reading internal warm surface.

The algorithm for compositing individual values of T_B to produce CELL-ALL is under development (12/79).

g. Quality Assessment

1. Validation - An extensive program of validation has been conducted. Based principally on satellite under-flights with a simulator installed on the NASA CV-990, publication of the data is planned for 1980.

2. DBMS Operations - To be determined.

3. Confidence Level/Accuracy - To be determined. Errors in IFOV location have been noted.

4. Usage Guidance - Data will be available from NSSDC for studies of correlation with geophysical parameters.

h. Output Products (In Archive)

Tapes archived and routinely delivered, or available for delivery, to users are:

1. TAT, Antenna Temperature Tape - Contains calibrated antenna temperatures and earth locations for each IFOV for each polarization. Also contains ephemeris, attitude, and housekeeping information.

2. MAP-30, Mapped Parameter of 37-GHz Channel - Contains north and south polar projections of 37-GHz brightness temperatures.

3. CELL-ALL - Contains horizontal and vertical polarization brightness temperatures and seasonal geographic filters for each of the five channels at 150-km resolution for all but the 4.6-cm channel at 97.5-km resolution, for all but the 4.6-cm and 2.8-cm channels at 60-km resolution, and for only the 0.8-cm channel at 30-km resolution. Data are grouped by cells and bands of various sizes, but each combination of cells and bands equals 780-km^2. Location coordinates are given for each cell and band.

Tapes archived, but not routinely delivered to users are:

1. UFO-S User Formatted Output Tape - SMMR - Contains SMMR housekeeping telemetry and primary sensor data.

2. ILT-S, Image Location Tape - SMMR - Contains spacecraft ephemeris, attitude, and GMT-to-spacecraft time correction sufficient to locate SMMR IFOV's, plus bit slip information.

3. MATRIX-30, Mapped Data Matrix of 37-GHz channel - Contains north and south polar projection map matrices of 37-GHz brightness temperatures. Each temperature is color coded. Contains additional reference and title information for completing color film spec display F231320.

As of December 1979, 6 months of TAT have been produced on an experimental basis. Six weeks of CELL-ALL tapes have been produced on an experimental basis. They are not yet available to NSSDC.

Image products - Polar stereographic maps of the 37-GHz brightness temperatures have been produced on an experimental basis. Each map covers a 6-day interval.

i. Availability

Experimental products have been delivered to the members of the SMMR Nimbus Experiment Teams for validation studies and algorithm development or refinement. These activities are expected to make data available through NSSDC late in 1980.

j. Access

1. Location - Not yet determined by NSSDC.

2. Procedure for Obtaining - Contact NSSDC.

k. Contacts

1. Data Producer

Dr. Per Gloersen, Sensor Scientist
Code 913
Goddard Space Flight Center

2. Data Manager

Mr. James R. Greaves
Code 910.2
Goddard Space Flight Center

l. References

1. Other data sets - See Sea Surface Temperature, Sea Ice.

2. Support Documentation - Specifications for each tape product are available from NSSDC when data are achived.

3. Related Literature:

The Nimbus 7 User's Guide. GSFC, 1978.

An Algorithm for Retrieval of Ocean Surface and Atmospheric Parameters from the Observations of the Scanning Multi-channel Microwave Radiometer (SMMR). T.T. Wilheit and A.T.C. Chang. GSFC TM 8027, May 1979.

m. Summary/Sample

1. Averages, Variances - (Not available - 12/79).

2. Histograms - (Not available, 12/79).

3. Sample Output -

a. Tape (To be provided).

b. Map.

APPENDIX D

SAMPLE INTRODUCTORY LETTER

November __, 1980

Refer to 311-BF:amc

Dear _____:

Thank you for agreeing to meet with me on _____ to discuss your applications of oceanographic data. The information obtained will directly impact the design of user-oriented, general-purpose software.

The purpose of the study is to define for NASA's Office of Space and Terrestrial Applications the functional requirements for a general-purpose data base management system (DBMS) for scientific and experimental data. Such a system would be made available to projects which collect, store, manipulate, catalog, and/or display various types of oceanographic data. These data may include sensor measurements, imagery, plots, text, and other data types derived from many possible sources, including satellites, aircraft, balloons, vessels, buoys, ground sensors, etc. The ultimate objective of this general-purpose DBMS is to minimize the redundant development of data management software and to facilitate the direct exchange of data and information between NASA and the oceanographic community.

As part of the study, several major data management systems have been analyzed and documented such as the Fleet Numerical Oceanography Center (FNOC), National Oceanographic Data Center (NODC), Satellite Data Services Division (SDSD), Seasat Project Data Processing System, Seasat SAR, JPL's Video Information Communication and Retrieval System (VICAR), and the Atmospheric and Oceanographic Information Processing System (AOIPS) that was developed by Goddard SFC.

You have been identified as a leading authority in the field of oceanography, and your participation in this effort will facilitate the identification of specific data base management functions and capabilities that should be developed in future systems. Through our discussion, I hope to better understand and detail your concerns about data management. The initial results of this task indicate that some of the general problems that oceanographic data users experience include finding the appropriate data source, receiving data in a short time frame, reading and formatting data, merging and comparing different data sets, and providing the data in a useful form for the particular application.

In order to structure the conversation and to indicate our particular interests, we have compiled some fairly general questions. Examples of the general questions, which may lead to more specific inquiries, are:

- How do you procure oceanographic data? What types of data?

- What are your current data processing facilities? Access capabilities? System software?

- What types of application programs are utilized in your research for formatting/interpreting the data? How often is it necessary to modify them? Why do you modify them?

- How do you manipulate the data? Do you find it necessary to merge different data sets? Compare or sort data sets? Data corrections? If so, how are these operations accomplished? Why are these operations necessary? How much time/manpower is consumed?

- What types of information about available data sets would be helpful for your purposes?

- What types of special tools could you utilize for your final products? (Report generation, archive, user data entry, security functions, for example).

Naturally, not all of these questions may be applicable to you nor are they necessarily exhaustive, but they indicate the breadth of our interest.

Thank you again for your assistance.

Sincerely,

Brad Fujimoto

APPENDIX E

ACRONYMS

ADBMS	Applications Data Base Management System
DBMS	Data Base Management System
DOE	U.S. Department of Energy
FNOC	Fleet Numerical Oceanography Center
GSFC	Goddard Space Flight Center
JPL	Jet Propulsion Laboratory
NOAA	National Oceanographic and Atmospheric Administration
NRT	Near-real-time
NWS	National Weather Service
ODSI	Ocean Data Systems, Inc.
ODUS	Ocean Data Utilization System
OPS	Oceanic Pilot System
OSTA	Office of Space and Terrestrial Applications (NASA)
SAR	Synthetic Aperture Radar
SASS	Seasat Satellite Scatterometer
SDDS	Satellite Data Distribution System
SOWM	Spectral Ocean Wave Model
SST	Sea Surface Temperature
TOPEX	Ocean Topography Experiment
Z	Zulu or Greenwich Mean Time